Kernels for Measures Defined on the Gram Matrix of their Support

Marco Cuturi
Princeton University
mcuturi@princeton.edu

September 7, 2009

Abstract

We present in this work a new family of kernels to compare positive measures on arbitrary spaces $\mathcal{X}$ endowed with a positive kernel κ, which translates naturally into kernels between histograms or clouds of points. We first cover the case where $\mathcal{X}$ is Euclidian, and focus on kernels which take into account the variance matrix of the mixture of two measures to compute their similarity. The kernels we define are semigroup kernels in the sense that they only use the sum of two measures to compare them, and spectral in the sense that they only use the eigenspectrum of the variance matrix of this mixture. We show that such a family of kernels has close bonds with the laplace transforms of nonnegative-valued functions defined on the cone of positive semidefinite matrices, and we present some closed formulas that can be derived as special cases of such integral expressions. By focusing further on functions which are invariant to the addition of a null eigenvalue to the spectrum of the variance matrix, we can define kernels between atomic measures on arbitrary spaces $\mathcal{X}$ endowed with a kernel κ by using directly the eigenvalues of the centered Gram matrix of the joined support of the compared measures. We provide explicit formulas suited for applications and present preliminary experiments to illustrate the interest of the approach.

1 Introduction

Defining meaningful kernels on positive measures is an important issue in the field of kernel methods, as it encompasses the topic of comparing histograms, bags-of-components or clouds of points, which all arise very frequently in applications dealing with structured data.

In the pioneering applications of support vector machines to structured data, histograms were often treated as simple vectors and used as such through the standard Gaussian or polynomial kernels [Joa02]. Yet, more adequate kernels which exploit the specificities of histograms have been proposed since. Namely,

the fact that histograms are vectors with nonnegative coordinates [HB05], and whose sum may be normalized to one, that is cast as discrete probability measures and treated under the light of information geometry [LL05, Leb06]. Since such histograms are usually defined on bins which are not equally dissimilar, as is for instance the case with color, words or amino-acid histograms, further kernels which may take into account an a priori inter-bin similarity where subsequently proposed [KJ03, CFV05, HB05] as an attempt to include with more accuracy a prior knowledge on the considered components, through the knowledge of a prior kernel κ for instance.

In this paper we investigate further such kind of kernels between two measures, which can conveniently describe the similarity between two clouds of points by only considering their Gram matrices. In this sense we reformulate and extend the results of [CFV05] whose framework we briefly recall:

The set $M_+^b(\mathcal{X})$ of bounded positive measures on a set $\mathcal{X}$ is a cone, and from a more elementary algebraic viewpoint a semigroup[1]. In that sense, a natural way to define kernels suited to the geometry of $M_+^b(\mathcal{X})$ is to study the family of semigroup functions on $M_+^b(\mathcal{X})$, as introduced in [BCR84], that is real-valued functions ψ defined on $M_+^b(\mathcal{X})$ such that the map $(\mu, \mu') \mapsto \psi(\mu + \mu')$ is either positive or negative definite. The Jensen-divergence, which is computed through the entropy of the mixture of two measures is such an example, as recalled in [HB05].

Given the complexity of evaluating entropies for finite samples, it is shown in [CFV05] that similar quantities can be defined for measures by only taking into account the variance matrix $\Sigma(\mu+\mu')$ of the mixture of two measures. This has two clear advantages. First, variances are easy to compute given atomic measures, that is measures with finite support, which are usually considered in most applications. Second, the eigenspectrum of the variance matrix of an (atomic) probability measure is known to be the same as, up to zero eigenvalues and an adequate centralization, the eigenspectrum of the dot-product matrix of the support of the same measure. This fact paves the way to consider kernels defined on Gram matrices rather than on variance matrices, regardless of the structure of $\mathcal{X}$, as was first hinted in [KJ03].

More precisely, the authors of [CFV05] first prove that for a variance matrix $\Sigma(\mu + \mu')$, the determinant $|\frac{1}{\eta}\Sigma(\mu + \mu') + I_n|^{-\frac{1}{2}}$ for $\eta > 0$ is a positive definite (p.d.) kernel between two measures μ, μ' on an Euclidian space $\mathcal{X}$ of dimension n. Second, they prove that this quantity can be cast into a reproducing kernel Hilbert space (rkHs) associated with a kernel κ on $\mathcal{X}$, regardless of the nature of $\mathcal{X}$, by using directly a centered Gram matrix $\mathcal{K}_{\mu,\mu'}$ of all elements contained in the support of both μ and μ'.

We are interested in this paper in characterizing other functions φ defined on matrices such that *(i)* $\mu, \mu' \mapsto \varphi(\Sigma(\mu+\mu'))$ is either positive or negative definite, *(ii)* φ is spectral[2] and *(iii)* φ is invariant to the addition of null eigenvalues,

[1] In this paper, a semigroup will be a non-empty set S endowed with a commutative addition, such that for $s, t \in S$, their sum $s + t = t + s \in S$, and a neutral element e such that $s + e = s$

[2] A function f defined on symmetric matrices is spectral, or orthogonally invariant, if for

that is, for two square p.d. matrices A, B which may not have the same size, $\varphi(A) = \varphi(B)$ if A and B have the same *positive* eigenvalues taken with their multiplicity, regardless of the multiplicity of 0 in their eigenspectrum. It is easy to check that both $|\frac{1}{\eta} \cdot + I|$ and the trace fulfill condition *(iii)*.

If φ satisfies condition *(i)* and *(ii)*, we call the composed function $\psi = \varphi \circ \Sigma$ a *semigroup spectral positive (resp. negative) definite* (s.s.p.d., resp. s.s.n.d.) function on $M_+^b(\mathcal{X})$. Note that the task of defining such functions ψ is *not* equivalent to defining directly positive or negative definite functions φ on the semigroup of p.d. matrices, since the underlying semigroup operation is the addition of measures and not that of the variance matrices of the measures, as recalled in Equation (1). When φ is further invariant to null eigenvalues *(iii)* ψ can be cast in Hilbert spaces of infinite dimensions to compare degenerated variance operators, which will be in the context of this paper an rkHs built on $\mathcal{X}$ through a kernel κ.

This paper is structured as follows: we introduce in Section 2 an alleviated formalism for semigroup functions, and propose a general link between s.s.p.d. functions and the Laplace transform of functions defined on matrices in Section 3. We review then in Section 4 different s.s.p.d. functions, notably a function which satisfies criteria *(iii)* and which does not requite any regularization. We provide explicit formulas and test the kernel derived from such a function on a benchmark classification task involving handwritten digits in Section 5.

2 Semigroup Functions on Bounded Subsets of $M_+^b(\mathcal{X})$

We consider $\mathcal{X}$, an Euclidian space of dimension n endowed with Lebesgue's measure and restrict $M_+^b(\mathcal{X})$ to measures with finite first and second moments. In such a case, the variance of a measure μ of $M_+^b(\mathcal{X})$ can be defined as:

$$\Sigma(\mu) = \mu[xx^\top] - \mu[x]\mu[x]^\top.$$

Writing $\bar{\mu}$ for $\mu[x]$, we recall an elementary result for two measures μ, μ' of $M_+^b(\mathcal{X})$,

$$\Sigma(\mu + \mu') = \Sigma(\mu) + \Sigma(\mu') - \left(\bar{\mu}\bar{\mu}'^\top + \bar{\mu}'\bar{\mu}^\top\right), \tag{1}$$

which highlights the nonlinearity of the variance mapping.

We write $\mathbf{P}_n$ for the cone of real, symmetric and positive semidefinite matrices, and $\mathbf{P}_n^+$ for its subset of (strictly) p.d. matrices. In this paper, the assumption that for a measure μ its variance $\Sigma(\mu)$ is in $\mathbf{P}_n$ is crucial for most calculations, and this is ensured for sub-probability measures, that is is measures μ such that $|\mu| = \mu(\mathcal{X}) \leq 1$, since we then have that

$$\Sigma(\mu) = \mu[(x - \bar{\mu})(x - \bar{\mu})^\top] + (1 - |\mu|)\,\bar{\mu}\bar{\mu}^\top \in \mathbf{P}_n. \tag{2}$$

any real $n \times n$ orthogonal matrix H, that is such that $HH^\top = I_n$, $f(HAH^\top) = f(A)$. In that case f only depends on the eigenspectrum of A. See [BL00]

Furthermore, we will also need the identity $\Sigma(\mu) = \mu[(x - \bar{\mu})\,(x - \bar{\mu})^\top]$ in order to make the link between the dot-product matrix of the support of μ and its variance matrix, which is why we restrict our study to probability measures $M_+^1(\mathcal{X})$. $M_+^1(\mathcal{X})$ is not, however, a semigroup, since it is not closed under addition, due to the constraint on $|\mu|$. To cope with this contradiction, that is to use semigroup-like functions of the type $(\mu, \mu') \to \psi(\mu + \mu')$ where ψ is only defined on a subset of the original semigroup, and where this subset may not be itself a semigroup, we define the following extension to the original definition of semigroup functions which, although technical, is also useful to recall the actual definitions of positive and negative definiteness for semigroup functions.

Definition 1 (Semigroup kernels on subsets) *Let $(S, +)$ be a semigroup and $U \subset S$ a nonempty subset of S. A function $\psi : U \to \mathbb{R}$ is a p.d. (resp. n.d.) semigroup function on U if*

$$\sum_{i,j} c_i c_j \psi(s_i + s_j) \geq 0 \;\; (\textit{resp} \leq 0)$$

holds for any $n \in \mathbb{N}$; any $s_1, \ldots, s_n \in S$ such that $s_i + s_j \in U$ for $1 \leq i \leq j \leq n$; and any $c_1 \ldots, c_n \in \mathbb{R}$ (resp. with the additional condition that $\sum_i c_i = 0$)

In practice, stating that a function ψ defined on the subset $M_+^1(\mathcal{X})$ is positive (resp. negative) definite is equivalent to stating that the kernel for two elements μ, μ' of $M_+^1(\mathcal{X})$ defined as

$$(\mu, \mu') \mapsto \psi\left(\frac{\mu + \mu'}{2}\right)$$

is positive (resp. negative) definite. Finally, we write $\Sigma^{-1}(\mu)$ for $(\Sigma(\mu))^{-1}$ when appropriate.

3 Laplace Transforms of Matrix Functions and s.s.p.d. functions

We show in this section how s.s.p.d. functions on $M_+^1(\mathcal{X})$ can be defined through the Laplace transform of a nonnegative-valued function defined on the cone $\mathbf{P}_n^+$, through the following lemma.

Lemma 2 *For any $S \in \mathbf{P}_n$, the real-valued function defined on $M_+^1(\mathcal{X})$,*

$$\mu \mapsto \langle \Sigma(\mu), S \rangle$$

is a negative definite semigroup function.

Proof. For any $k \in \mathbb{N}$, any $c_1, \ldots, c_k \in \mathbb{R}$ such that $\sum_i c_i = 0$ and any $\mu_1, \ldots, \mu_k \in M_+^1(\mathcal{X})$ such that $\mu_i + \mu_j \in M_+^1(\mathcal{X})$, we have using Equation (1)

that

$$\sum_{i,j} c_i c_j \left\langle \Sigma\left(\mu_i + \mu_j\right),\, S \right\rangle = \left\langle \sum_{i,j} c_i c_j \Big(\Sigma(\mu_i) + \Sigma(\mu_j) - \left(\bar{\mu}_i \bar{\mu}_j^\top + \bar{\mu}_j \bar{\mu}_i^\top\right) \Big),\, S \right\rangle$$

$$= -\left\langle \sum_{i,j} c_i c_j \left(\bar{\mu}_i \bar{\mu}_j^\top + \bar{\mu}_j \bar{\mu}_i^\top\right),\, S \right\rangle = -2 \sum_{i,j} c_i c_j \bar{\mu}_i^\top S \bar{\mu}_j \leq 0. \blacksquare$$

Note that this function is actually n.d. for all measures of $M_+^b(\mathcal{X})$, regardless of their total weight $|\mu|$. The case $S = I_n$ yields the simple function $\psi_{\mathrm{tr}} \stackrel{\mathrm{def}}{=} \mu \mapsto \mathrm{tr}\,\Sigma(\mu)$, which provides interesting results in practice, and boils down to a fast kernel on clouds of points, which we will review briefly in Section 5. For any nonnegative-valued function $f : \mathbf{P}_n^+ \to \mathbb{R}^+$ defined on the cone of p.d. matrices, we write

$$\mathcal{L}f(Z) = \int_{S \in \mathbf{P}_n^+} e^{-<Z,S>} f(S) dS \tag{3}$$

for the Laplace transform of f evaluated in $Z \in \mathbf{P}_n^+$, when the integral exists.

Proposition 3 *For any spectral function $f : \mathbf{P}_n^+ \to \mathbb{R}^+$, the mapping*

$$\mu \mapsto \mathcal{L}f\left(\Sigma(\mu)\right)$$

defined for all measures $\mu \in M_+^1(\mathcal{X})$ such that $\Sigma(\mu) \in \mathbf{P}_n^+$ is a s.s.p.d. function.

Proof. The integral when it exists is a sum of p.d. semigroup functions through Schoenberg's theorem [BCR84, Theorem 3.2.2], and is hence p.d.$\blacksquare$

Laplace transforms of functions defined on matrices is an extensive subject and we refer to [Mat93, Section 4] for a short survey. In the case where $f = 1$ we recover the characteristic function of the cone $\mathbf{P}_n^+$, and its logarithm, $\ln \mathcal{L}1(A) = C - \frac{n+1}{2} \log|A|$, is known as the universal barrier [Gül96] of the cone $\mathbf{P}_n^+$, with numerous applications in convex optimization.

We recall now a well-known result of multivariate analysis based on zonal polynomials (see [Tak84, MPH95] for an exhaustive presentation of these), which may not, however, be of immediate use for an application in kernel methods. To be short, zonal polynomials $C_\alpha(A)$ are polynomials in the eigenvalues of a matrix A with positive coefficients [MPH95, Remark 4.3.6], and thus nonnegative-valued spectral functions, indexed by the partitions α of an integer a. Namely, for $a \in \mathbb{N}$, we write $\alpha = (a_1, a_2, \ldots, a_n)$ for a partition of a into not more than n parts, that is $a_1 + a_2 + \cdots + a_n = a$ and $a_1 \geq a_2 \geq \cdots \geq a_n$. The following result follows from [MPH95, Theorem 4.4.1] where we have dropped constants which only depend of n and α for more readability:

Corollary 4 *Given $a \in \mathbb{N}$ and a partition α of a, the real-valued zonal kernel ψ_α is a s.s.p.d. function on $M_+^1(\mathcal{X})$, with*

$$\psi_\alpha : \mu \mapsto \frac{C_\alpha(\Sigma^{-1}(\mu))}{|\Sigma(\mu)|^{-\frac{1}{2}n}},$$

through the identity $\int_{S\in\mathbf{P}_n^+} e^{-<\Sigma,S>}|S|^{t-\frac{1}{2}(n+1)}C_\alpha(S)dS \propto |\Sigma|^{-t}C_\alpha(\Sigma^{-1})$*, for* $t > \frac{1}{2}(n-1)$.

Actual expressions for zonal polynomials of order $a \leq 10$ are currently known, and the use of Wishart densities for f can be seen as a special case of such evaluations. It is also clear that finite and infinite linear combinations of such zonal kernels, with the speculation that they might be a useful basis for a subcategory of s.s.p.d. functions, can be carried out in the spirit of equations provided in [MPH95, Lemmas 4.4.5&6] and yield convenient formulas, such as

$$\sum_{a=k}^{\infty}\sum_{\alpha}\psi_\alpha : \mu \mapsto \frac{e^{\mathrm{tr}\,\Sigma^{-1}(\mu)}(\mathrm{tr}\,\Sigma^{-1}(\mu))^k}{|\Sigma(\mu)|^{-\frac{1}{2}n}},$$

which is a s.s.p.d. function for any $k \geq 0$. However, the weak point of these expressions when used in our setting is that they tend to be extremely degenerated when the eigenspectrum of Σ vanishes, due to the high power of the denominator and to the fact that the eigenspectrum of Σ^{-1}, not Σ, is considered implicitly. Hence, we do not see at the moment how one would obtain expressions satisfying condition *(iii)*, even through the use of regularization. To handle this problem, we focus in the next section on degenerated integrations, that is we consider an extension of the Laplace transform setting defined in Equation (3) to degenerated functions f defined on families of semidefinite matrices of $\mathbf{P}_n$.

4 Degenerated Integrations on Semidefinite Matrices of Rank 1

We restrict the integration domain to only consider the subspace of $\mathbf{P}_n$ of matrices of rank 1, that is matrices of the form $yy^\top$ where $y \in \mathbb{R}^n$. The Euclidian norm $y^\top y$ of y is the only positive eigenvalue of $yy^\top$ when $y \neq 0$, hence only real-valued functions of $y^\top y$ can be spectral. Following the proof of Proposition 3, and for any nonnegative-valued function $g : \mathbb{R}^+ \to \mathbb{R}^+$, we observe thus that

$$\psi_g : \mu \mapsto \int_{\mathbb{R}^n} e^{-y^t\Sigma(\mu)y}g(y^\top y)dy \tag{4}$$

is a s.s.p.d. function on $M_+^1(\mathcal{X})$, noting simply that $\mathrm{tr}(\Sigma(\mu)yy^\top) = y^t\Sigma(\mu)y$. We start our analysis with a simple example for g, which can be computed in close form.

4.1 The case $g : x \mapsto x^i$

For a matrix $A \in \mathbf{P}_n^+$ such that $\mathrm{mspec}\,A = \{\lambda_1, \ldots, \lambda_n\}$, we set $\gamma_0(A) = 1$ and write for $1 \leq i \leq n$,

$$\gamma_i(A) \stackrel{\mathrm{def}}{=} \sum_{|j|=i} \frac{\prod_{k=1}^n \Gamma(j_k + \frac{1}{2})}{\lambda_1^{j_1}\cdots\lambda_n^{j_n}}$$

where the summation is taken over all families $j \in \mathbb{N}^n$ such that the sum of their elements $|j|$ is equal to i. Writing σ_n for $(2\pi^{-\frac{1}{2}})^{\frac{n}{2}}$ we have with these notations that for all $1 \leq i \leq n$,

Corollary 5 *The function $\psi_i : \mu \mapsto \sigma_n \sqrt{\gamma_n} \cdot \gamma_i(\Sigma(\mu))$ is a s.s.p.d. functions on $M_+^1(\mathcal{X})$.*

Proof. Let $\mu \in M_+^1(\mathcal{X})$, and write $\operatorname{mspec} \Sigma(\mu) = \{\lambda_1, \ldots, \lambda_n\}$. Then by an appropriate base change we have for $g_i : x \mapsto x^i$, $i \leq n$,

$$\psi_{g_i}(\mu) = \int_{\mathbb{R}^n} e^{-\sum_{k=1}^n \lambda_k y_k^2} \left(\sum_{k=1}^n y_k^2 \right)^i dy = \int_{\mathbb{R}^n} e^{-\sum_{k=1}^n \lambda_k y_k^2} \sum_{|j|=i} \prod_{k=1}^n y_k^{2j_k} dy$$

$$= \sum_{|j|=i} \prod_{k=1}^n \int_{\mathbb{R}} e^{-\lambda_k y_k^2} y_k^{2j_k} dy_k = \sum_{|j|=i} \prod_{k=1}^n \Gamma(j_k + \frac{1}{2}) \lambda_k^{-j_k - \frac{1}{2}} = \sigma_n \sqrt{\gamma_n} \cdot \gamma_i(\Sigma(\mu)). \blacksquare$$

The inverse generalized variance is recovered as ψ_0. We refer now to Lancaster's formulas [Ber05, p.320] to express more explicitly the cases $i = 1, 2, 3$, where we write Σ for $\Sigma(\mu)$:

$$\psi_1(\mu) = \frac{\sigma_n}{\sqrt{|\Sigma|}} \left[\operatorname{tr} \Sigma^{-1} \right], \quad \psi_2(\mu) = \frac{\sigma_n}{\sqrt{|\Sigma|}} \left[(\operatorname{tr} \Sigma^{-1})^2 + 2 \operatorname{tr} \Sigma^{-2} \right],$$
$$\psi_3(\mu) = \frac{\sigma_n}{\sqrt{|\Sigma|}} \left[(\operatorname{tr} \Sigma^{-1})^3 + 6(\operatorname{tr} \Sigma^{-1})(\operatorname{tr} \Sigma^{-2}) + 8 \operatorname{tr} \Sigma^{-3} \right].$$

Although the functions ψ_i are s.s.p.d., they are mainly defined by the lowest eigenvalues of $\Sigma(\mu)$. These functions can all be regularized, by adding a weighted identity matrix I_n to Σ, while still preserving their positive definiteness as can be easily justified by using the functions $g_i(x) = e^{-x} x^i$ to penalize for large values of $y^\top y$. In such a case however, and to the notable exception of ψ_0, this regularization prevents the above functions to be invariant to the addition of a zero eigenvalue to the spectrum of $\Sigma(\mu)$. Intuitively, this degeneracy is due to the fact that we integrate on the whole or R^n, notably on $\ker \Sigma(\mu)$, where the contribution of $\exp(-y^\top \Sigma(\mu) y)$ is infinite. We propose to solve this issue by considering more specifically the contribution of each sphere $\{y | y^\top y = t\}$ to the overall summation in the case where $g = 1$.

4.2 The case $g : x \mapsto \delta_t$ and its variants

The question of integrating $\exp(-y^\top \Sigma y)$ over compact balls $\{y \in \mathbb{R}^n \,|\, y^\top y \leq t\}$ or spheres $\{y \in \mathbb{R}^n \,|\, y^\top y = t\}$ is closely related to the evaluation of the distribution of quadratic forms in normal variates [MP92]. Given a matrix $Q \in \mathbf{P}_n$ and a random vector y in $\mathbb{R}^n$ following a normal law $\mathcal{N}(m, V)$ with $V \in \mathbf{P}_n^+$, the density $h[Q, V, m]$ of the values of $y^\top Q y$, that is

$$h[Q, V, m](t) dt = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \int_{t < y^\top Q y < t + dt} e^{-(y-m)^\top V^{-1} (y-m)} \, dy, \quad (5)$$

can be evaluated explicitly in terms of the eigenvalues of $V^{\frac{1}{2}}QV^{\frac{1}{2}}$ through series expansions [MP92], as well as the distribution of h and its Laplace transform. We note that these expressions are similar to that of the elementary contribution of the sphere $\{y^\top y = t\}$ when $g = 1$,

$$f_\mu(t)dt = \int_{t<y^\top y<t+dt} e^{-y^\top \Sigma(\mu) y}\, dy. \tag{6}$$

The difference between expressions (6) and (5) is that the variance $\Sigma^{-1}(\mu)$ may be potentially infinite if Equation (6) is directly translated in terms of Equation (5), while the normalization term in $\sqrt{|\Sigma(\mu)|}$ does not appear in the s.s.p.d. function of Equation (6). It turns out that these two problems can be easily canceled out. We consider the Laplace transform of $f_\mu(t)$ defined as $L_\mu(s) = \int_{t\geq 0} e^{-st} f_\mu(t)dt$, which is trivially a s.s.p.d. as a sum of elementary s.s.p.d. functions.

Lemma 6 *For* $\operatorname{mspec}\Sigma(\mu) = \{\lambda_1, \ldots, \lambda_n\}$*, define the sequences* $d_k = \frac{1}{2}\sum_{j=1}^n \lambda_j^k$ *for* $k \geq 1$*,* $c_0 = 1$ *and* $c_k = \frac{1}{k}\sum_{r=0}^{k-1} d_{k-r}c_r$ *for* $k \geq 1$*. Then,*

$$f_\mu(t) = \pi^{\frac{n}{2}} \sum_{k=0}^{\infty} (-1)^k c_k \frac{t^{\frac{n}{2}+k-1}}{\Gamma(\frac{n}{2}+k)}; \quad L_\mu(s) = \pi^{\frac{n}{2}} \sum_{k=0}^{\infty} (-1)^k c_k\, s^{-\frac{n}{2}+k}.$$

Proof. Take the exact value computation for

$$\frac{\pi^{\frac{n}{2}}}{|\Sigma(\mu) + 2\eta I_n|^{\frac{1}{2}}} h[I_n, (\frac{1}{2}\Sigma(\mu) + \eta I_n)^{-1}, 0](t) = \int_{y^\top y = t} e^{-y^\top(\Sigma(\mu)+\eta I_n)y}\, dy.$$

and let $\eta \to 0$ to obtain through [MP92, Lemma 4.2b.2 & Theorem 4.2b.1] the corresponding formulas, both for f_μ and L_μ■

Proposition 7 *By considering* $s = 1$ *in* $L_\mu(s)$ *from Lemma 6, and noting that* $d_k = \frac{1}{2}\operatorname{tr}(\Sigma(\mu)^k)$ *we have that*

$$\mu \mapsto \psi_M(\mu) = \sum_{k=0}^{\infty} (-1)^k c_k$$

is a s.s.p.d. function on $M_+^1(\mathcal{X})$*, invariant to the addition of null eigenvalues, defined when the spectrum of* $\Sigma(\mu)$ *is strictly upper-bounded by* 1*.*

Although all terms d_k are invariant by the addition of a null eigenvalue, f_μ cannot fulfill condition (iii) because of the numerator in $\Gamma(\frac{n}{2}+k)$ which depends on both the dimension n and the summation variable k.

We refer to the proof [CFV05, Theorem 7] to show that for an atomic measure μ such that $|\mu| = 1$, that is $\mu \in \mathrm{Mol}_+^1(\mathcal{X})$, $\psi_M(\mu)$ can be either expressed in terms of the spectrum of its variance matrix or in the spectrum of its centered dot-product matrix. Thus, if $\mathcal{X}$ is now an arbitrary space endowed with a kernel κ, the centered Gram matrix corresponding to the support of μ can be

used directly as an input for a s.s.p.d. function such as ψ_M in order to define a p.d. kernel on $\mathrm{Mol}_+^1(\mathcal{X})$. We detail in Section 5 the formulation of such a kernel, along with two other examples.

Finally, as a closing remark for this section, note that the representation proposed in Proposition 3 is not exhaustive to our knowledge, and should not be confused with the integral representation of semigroup p.d. functions as sums of semicharacters, studied in [CFV05]. First, the functions $\mu \mapsto e^{-\langle \Sigma(\mu), S\rangle}$ are not semicharacters[3] of the semigroup $M_+^1(\mathcal{X})$, since

$$e^{-\langle \Sigma(\mu+\mu'), S\rangle} \neq e^{-\langle \Sigma(\mu)+\Sigma(\mu'), S\rangle}$$

in the general case. Second, the class of functions considered through Lemma 2 is far from characterizing all semigroup n.d. functions on $M_+^1(\mathcal{X})$ since, through [BCR84, Corrolary 3.2.10, p.78], we have that for any $S \in \mathbf{P}_n$ and $0 < \beta < 1$ both

$$\mu \mapsto \langle \Sigma(\mu), S\rangle^\beta, \quad \text{and } \mu \mapsto \ln(1 + \langle \Sigma(\mu), S\rangle),$$

are semigroup n.d. functions. Note that if we use for $y \in \mathbb{R}^n$ and $m \geq 1$ a n.d. function of the type

$$\mu \mapsto \frac{m+n}{2} \ln\left(1 + \frac{1}{m} y^\top \Sigma(\mu) y\right),$$

and exponentiate it in the spirit of Equation (4), we recover the integration of the Student multivariate distribution for vectors of $\mathbb{R}^n$, which boils down again to a kernel that is proportional to ψ_0.

5 Explicit Formulas for Atomic Measures and Experiments

Given two clouds of weighted points $\gamma = (x_i, a_i)_{i=1}^d$ and $\gamma' = (y_i, b_i)_{i=1}^{d'}$, we show how to compute three different kernels which satisfy condition *(iii)*, namely ψ_{tr}, ψ_0 and ψ_M, and compare them by studying their performance on a multiclass classification task.

5.1 Formulations for Clouds of Points

The mixture of γ and γ' can be expressed as $\gamma'' = \{(x_i, \frac{1}{2}a_i)_{i=1}^d, (y_j, \frac{1}{2}b_j)_{j=1}^{d'}\}$. By writing $d'' = d + d'$ and

$$\mathcal{K}_\gamma = [\kappa(x_i, x_j)]_{i,j \leq d}, \quad \mathcal{K}_\gamma = [\kappa(y_i, y_j)]_{i,j \leq d'}, \text{and } \mathcal{K}_* = [\kappa(x_i, y_j)]_{i \leq d,\, j \leq d'}$$

we can further express the $d'' \times d''$ Gram matrix of the mixture γ'' as

$$\mathcal{K}_{\gamma''} = \begin{pmatrix} \mathcal{K}_\gamma & \mathcal{K}_* \\ \mathcal{K}_*^\top & \mathcal{K}_{\gamma'} \end{pmatrix}.$$

[3] Semicharacters of a semigroup S are, following the definition of [BCR84], real-valued functions ρ on S such that for $s, t \in S$, $\rho(s+t) = \rho(s)\rho(t)$

As can be seen in [CFV05], the spectrum of the Gram matrices cannot be taken as such since they require a centralization of the form

$$\tilde{\mathcal{K}}_{\gamma''} = (I_{d''} - \mathbb{1}_{d'',d''}\Delta_{\gamma''})\mathcal{K}_{\gamma''}(I_{d''} - \Delta_{\gamma''}\mathbb{1}_{d'',d''})\Delta_{\gamma''}, \tag{7}$$

with $\Delta_{\gamma''} = \mathrm{diag}(\frac{1}{2}a_i, \frac{1}{2}b_j)$ and $\mathbb{1}_{d'',d''}$ is the $d'' \times d''$ matrix of ones. The explicit formulas for the considered kernels, which we do not normalize and propose with bandwidth parameters are thus

$$k_{\mathrm{tr}}(\gamma,\gamma') = e^{-\frac{1}{t}\,\mathrm{tr}\,\tilde{\mathcal{K}}_{\gamma''}}$$

$$k_0(\gamma,\gamma') = |\frac{1}{\eta}\tilde{\mathcal{K}}_{\gamma''} + I_{d''}|^{-\frac{1}{2}}$$

$$k_M(\gamma,\gamma') = \sum_{k=0}^{\infty}(-1)^k c_k\,, \quad \text{with } c_0 = 1, c_k = \frac{1}{k}\sum_{r=0}^{k-1} d_{k-r}c_r \text{ and } d_k = \frac{1}{2}\,\mathrm{tr}([\delta\tilde{\mathcal{K}}_{\gamma''}]^k)$$

with $t > 0, \eta > 0$ and $\delta > 0$ such that $\delta < 1/\rho(\tilde{\mathcal{K}}_{\gamma''})$, where for a matrix $A \in \mathbf{P}_n$ such that $\mathrm{mspec}\,A = \{\lambda_1, \ldots, \lambda_n\}$, $\rho(A)$ is the spectral radius of A, that is $\max_{1\leq i\leq n}\lambda_i$.

We discuss now possible values for δ which will ensure that $\delta < 1/\rho(\tilde{\mathcal{K}}_\gamma)$ for any cloud-of-points γ and any kernel κ upper-bounded by one, that is $sup_{x\in\mathcal{X}}|\kappa(x,x)| \leq 1$. Through Equation (7), one can obtain that for any cloud of points $\gamma = (x_i, a_i)_{i=1}^d$,

$$\rho(\tilde{\mathcal{K}}_\gamma) \leq [\max(d \cdot a_{\max} - 1, 1)]^2 d \cdot a_{\max}$$

where we write $a_{\max}$ for the maximal weight of γ and we have bounded $\rho(\mathcal{K}_\gamma)$ by d, which corresponds to the case $\mathcal{K}_\gamma = \mathbb{1}_{d,d}$. Thus, any factor δ chosen so that

$$\delta < \frac{1}{[\max(d \cdot \omega - 1, 1)]^2 d \cdot \omega}$$

can be used to compare families of clouds of points whose maximal weights do not exceed ω and maximal size does not exceed $\frac{1}{2}d$. In the case where these clouds are bounded between $d_{\min}$ points (with weight $1/d_{\min}$) and $d_{\max}$ points, this condition is ensured for $\delta \leq (\frac{d_{\min}}{d_{\max}})^3$, which is far from being optimal in practical cases since the values of κ are more likely to be better distributed in the [0,1] range. This shows however that if we compare clouds of similar size δ can be equal to 1, and possibly above depending on the kernel κ which is used. We leave for future work the study of the convergence of the series $\sum_{k=0}^N(-1)^k c_k$ corresponding to the evaluation of k_M, although we note that in the practice of our experiments very few iterations (that is N set between 10 and 20) are sufficient to converge to the limit value, which reduces considerably the overall computation cost with respect to a straightforward eigenvalue decomposition of $\tilde{\mathcal{K}}_{\gamma''}$. Indeed, as is the case with the inverse generalized variance, this would have a cost of the order of d^3 while N computations of the traces $\mathrm{tr}[\delta\tilde{\mathcal{K}}_{\gamma''}]^k$ only grow in complexity Nd^2. It would be wise, however, to let N depend adaptively on the convergence of $\mathrm{tr}([\delta\tilde{\mathcal{K}}_{\gamma''}]^k)$ to 0, which is very much conditioned by the observed spectrum for κ.

5.2 Experiments on MNIST handwritten digits

We use the Experimental setting of [KJ03], also used in [CFV05] to compare the three previous kernels, namely, we sample 1.000 images from the MNIST database, that is 100 images per digit, and sample randomly clouds-of-pixels to compare such digits using the three kernels described above. The images, which are actually 28×28 matrices, are considered as clouds-of-pixels in the $[0,1]^2$ square, and we use a Gaussian kernel of width $\sigma = 0.1$ to evaluate the similarity between two pixels through κ, and use a three fold cross validation with five repeats to evaluate the performances of the kernels. The preliminary results shown in Table 1 show that the kernel ψ_M is competitive with both ψ_{tr} and the inverse generalized variance, which was itself shown to be effective with respect to other kernels in [CFV05], such as simple polynomial and Gaussian kernels.

Sample Size	$\psi_0, \eta = 0.01$	$\psi_{\mathrm{tr}}, t = 0.1$	$\psi_M, \delta = 1$
40 pixels	16.2	28.6	20.62
50 "	14.7	16.47	15.84
60 "	14.5	14.97	13.52
70 "	13.1	11.3	13
80 "	12.8	10.8	12.4

Table 1: Misclassification rate expressed in percents for the 3 s.s.p.d. functions used on a benchmark test of recognizing digits images, with 40 to 80 black points sampled from the original images.

References

[BCR84] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

[Ber05] Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. Princeton University Press, 2005.

[BL00] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*. Springer, New York, 2000.

[CFV05] Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *JMLR*, 6:1169–1198, 2005.

[Gül96] O. Güler. Barrier functions in interior point methods. *Mathematics of Operations Research*, 21:860–885, 1996.

[HB05] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proceedings of AISTATS 2005*, January 2005.

[Joa02] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publishers, 2002.

[KJ03] Risi Kondor and Tony Jebara. A kernel between sets of vectors. In T. Faucett and N. Mishra, editors, *Proc. of ICML'03*, pages 361–368, 2003.

[Leb06] Guy Lebanon. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):497–508, 2006.

[LL05] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *JMLR*, 6:129–163, January 2005.

[Mat93] Arak M. Mathai. *A Handbook of Generalized Special Functions for Statistical and Physical Sciences*. Oxford Science Publications, 1993.

[MP92] Arak M. Mathai and Serge B. Provost. *Quadratic Forms in Random Variables*. Number 126 in Statistics: Textbooks and Monographs. Dekker, 1992.

[MPH95] Arak M. Mathai, Serge B. Provost, and Takesi Hayakawa. *Bilinear Forms and Zonal Polynomials*. Number 102 in LNS. Springer Verlag, 1995.

[Tak84] Akimichi Takemura. *Zonal Polynomials*. Inst. Math. Stat. Lecture Notes, 1984.